

--	--	--	--	--	--	--	--	--	--

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2016/2017

TMA 7021 – DATA MINING AND ANALYTICS
(All sections / Groups)

10 OCTOBER 2016
8.00 p.m – 10.00 p.m
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This question paper consists of 6 pages with FOUR questions only.
2. Answer **ALL** questions.
3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

- (a) Differentiate between *predictive* data mining and *descriptive* data mining. List at least one technique for *predictive* data mining and *descriptive* data mining.

[2 marks]

- (b) Differentiate *eager* learners and *lazy* learners in supervised data mining. Provide an example for each learner.

[3 marks]

- (c) Define outlier and name the different approaches for outlier analysis.

[3 marks]

- (d) The information below presents the summary of “Income” of a particular company. Discuss why anomaly exists in the information given below.

```
> summary(Income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -8.7   14.6   35.0   53.5   67.0  615.0
```

[2 marks]

Continued...

QUESTION 2

(a) Study the following dataset carefully and write your answer in **R** codes.

Data Frame name = **dt_Student**

Name	Age	CGPA	Subject	Marks
Ting	20	3.5	Geography	95
Ting	20	3.5	History	90
Chong	25	3.2	Geography	89
Fatimah	18	3.1	Science	88
Fatimah	18	3.1	Technology	98
David	22	2.6	Science	98

- (i) Based on the dataset above, what data mining technique can be used to predict the Marks based on Age?

[1 mark]

- (ii) Using the `dplyr` package, calculate the average mark for each student and sort the records ascendingly according to *Name*.

[2 mark]

- (iii) Extend your solution in Question 2(a)(ii) by introducing a new column named "Achievement" that captures the following condition:

If average marks > 93, then "good" else "moderate"

[2 mark]

Continued...

(b) Study the table below.

Location ID	laundry	clinic	Fruit shop	restaurant	school
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

- (i) Calculate the *support* for the following item-set:
 $\{\text{restaurant}, \text{school}\}$
- (ii) Calculate the *confidence* for the following rule:
 $\{\text{fruit shop}, \text{clinic}\} \rightarrow \{\text{laundry}\}$
- (iii) Calculate the *Lift* for the following association rules:
 $\{\text{laundry}, \text{clinic}\} \rightarrow \{\text{fruit shop}\}$

[3 marks]

(c) Below is the output of association rule mining using R. Identify which rule is a redundant rule for another rule. Explain your finding.

lhs	rhs	support	confidence	lift
{Class=Crew, Sex=Male}	=> {Age=Adult}	0.3916402	1.0000000	1.0521033
{Class=Crew, Age=Adult}	=> {Sex=Male}	0.3916402	0.9740113	1.2384742
{Sex=Male, Survived=No}	=> {Age=Adult}	0.6038164	0.9743402	1.0251065
{Age=Adult, Survived=No}	=> {Sex=Male}	0.6038164	0.9242003	1.1751385
{Class=3rd, Sex=Male, Survived=No}	=> {Age=Adult}	0.1758292	0.9170616	0.9648435
{Class=3rd, Age=Adult, Survived=No}	=> {Sex=Male}	0.1758292	0.8130252	1.0337773
{Class=3rd, Sex=Male, Age=Adult}	=> {Survived=No}	0.1758292	0.8376623	1.2373791
{Class=Crew, Sex=Male, Survived=No}	=> {Age=Adult}	0.3044071	1.0000000	1.0521033
{Class=Crew, Age=Adult, Survived=No}	=> {Sex=Male}	0.3044071	0.9955423	1.2658514

[2 marks]

Continued...

QUESTION 4

Modeling political risk of a country is difficult. The difficulty stemmed mainly due to the uncertainty inherent in identifying the variables and the causal relationships between the variables. Assuming that you are given the following information:

Political risk is influenced by (i) *Social Stability*, (ii) *Government Stability*, (iii) *Business Environment*, and (iv) *Economic Policy*. *Government Stability* can be inferred from evidence about *war*, *income status*, *expropriation*, *terrorism*, *election*. *Business Environment* can be inferred from evidence about *recession*.

(a) Create a Bayesian Network based on the information above.

[5 marks]

(b) List any two *Observable (evidential)* nodes and two *Unobservable* nodes.

[2 marks]

(c) The *amount of information needed* to decide if an arbitrary example in S belongs to p or n is defined as

$$I(p, n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$

[3 marks]

Study the table below and calculate $I(p, n)$ for the class label.

RID	Age	Income	Student	Credit_rating	Class (buy_computer)
1	<=30	High	No	Fair	No
2	<=30	High	No	Excellent	No
3	31...40	High	No	Fair	Yes
4	>40	Medium	No	Fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	Low	Yes	Excellent	No
7	31...40	Low	Yes	Excellent	Yes
8	<=30	Medium	No	Fair	No
9	<=30	Low	Yes	Fair	Yes
10	>40	Medium	Yes	Fair	Yes
11	<=30	Medium	Yes	Excellent	Yes
12	31...40	Medium	No	Excellent	Yes
13	31...40	High	Yes	Fair	Yes
14	>40	Medium	No	Excellent	No

End of pages.